



Reductive, Exclusionary, Normalising: The Limits of Generative AI Music

RESEARCH ARTICLE

FABIO MORREALE 

MARCO A. MARTINEZ-RAMIREZ

RAUL MASU

WEIHSIANG LIAO

YUKI MITSUFUJI

**Author affiliations can be found in the back matter of this article*

]u[ubiquity press

ABSTRACT

Up until recently, most approaches to music generation were based on deductive logic: generative rules were devised on the basis of musicians' preferences, subjective appreciation and dominant music theories. Machine learning (ML) introduced a paradigm shift: vast datasets of existing music are used to train neural networks capable of generating new compositions supposedly without embedding predefined musical rules. We first outline how rule-based systems depend on a series of reductionist processes and assumptions about music that affect what can be generated. We then examine ML-based generative music systems and show that they are still unable to generate the full theoretical space of musical possibilities, they are still grounded on reductionist processes and their soundness is still affected by unquestioned assumptions. We also identify the limitations of semantic bridges used to form musical meaning and the epistemic framework of cascading modules. Finally, we propose that the artistic potential of ML systems might lie beyond attempts to replicate human music-making methods.

CORRESPONDING AUTHOR:

Fabio Morreale

Sony AI, Barcelona, Spain

fabio.morreale@sony.com

KEYWORDS:

Music ontology, music epistemology, diffusion, transformer, LLMs, Simondon, grammatisation, hyperreality

TO CITE THIS ARTICLE:

Morreale, F., Martinez-Ramirez, M.A., Masu, R., Liao, W., & Mitsufuji, Y. (2025). Reductive, Exclusionary, Normalising: The Limits of Generative AI Music. *Transactions of the International Society for Music Information Retrieval*, 8(1), 300-312. DOI: <https://doi.org/10.5334/tismir.256>

1 INTRODUCTION

Since the 1950s, computers have been used for a variety of music-generation tasks, employing different approaches – namely rule-based, stochastic and evolutionary systems (Griffith and Todd, 1999). Independent of the specific implementations, these approaches employ a deductive top-down logic of music generation, in which the musician/developer identifies the musical parameters to automate, formulates the generative rules and procedural logic and delegates their execution to the computer to generate music. While this approach has produced significant musical works, it inherently and unavoidably involves reductionist processes. These processes i) embed the generated music with the epistemological and ontological assumptions of the musician and the (scientific) episteme, ii) constrain the scope of music and its potentialities to a narrowly defined scope and iii) force music ontology (what is music) to be defined by the epistemic tools that are used to create it. Consequently, this approach yields only a partial and highly constrained account of music.

In recent years, several machine learning (ML)-powered generative music systems (GenAI) have been released, which are based on an inductive, bottom-up logic of generation (Pasquinelli, 2023). Unlike rule-based systems, GenAI systems¹ are not taught compositional rules that are grounded in arbitrarily defined grammar and do not encode specific (and inherently partial) understandings of music. Instead, they automatically capture musical features from massive quantities of music data, supposedly without directly referencing specific musical elements. As they do not require being taught explicit musical rules, GenAI might thus overcome the reductions and assumptions that affect the deductive logic of music generation.

While GenAI might depart from the specific sacrificial epistemic processes of rule-based systems, it remains an open question whether other forms of such processes persist. Are GenAI systems still imbued with implicit assumptions about music? Do they require human guidance and constraints to function effectively? Are they still subject to reductionist logic? Do they move beyond the effect of epistemology on music ontology, or do they still foreclose musical possibilities? While departing from the explicit rule-based logic of earlier systems, GenAI might remain structured by mechanisms that function as forms of implicit reductions. Whether these reductions constrain musical possibilities differently or more subtly than their deductive counterparts is a central concern of this paper.

While reduction has always played a crucial role in the formalisation and transmission of music, from notation systems to analytical frameworks, we argue that its effects in computational systems acquire a different weight. In these contexts, reductionist assumptions are

often invisibly embedded in models, treated as neutral abstractions and rarely subjected to epistemic scrutiny. It is precisely this lack of scrutiny that warrants the critical investigation we undertake in this paper. This work must be read alongside more technical contributions to music generation in music information retrieval (MIR). While often excelling in empirical capabilities, such technical work rarely offers sustained reflection on the epistemic and ontological assumptions they embed. By foregrounding these assumptions, our analysis complements technical perspectives and aims to broaden the discourse on generative systems in MIR.

Notably, in this article, we focus on end-to-end generative music systems: models capable of producing complete musical outputs directly from input data without intermediate human intervention, other than textual prompting. We acknowledge that other generative paradigms exist, including systems designed for continuation, inpainting or interactive co-creation. However, the scope of this paper is kept intentionally narrow for two reasons. First, end-to-end systems represent the most commercially relevant and publicly visible branch of current GenAI in music and therefore have the greatest impact on industry narratives and user expectations. Second, their design foregrounds the epistemic and ontological issues we investigate, since their training and inference processes encompass the full chain from representation to generation.

The remainder of the paper is structured as follows. [Section 2](#) examines rule-based (deductive) approaches to music generation, while [Section 3](#) focusses on representative ML (inductive) approaches. [Section 3](#) is deliberately more detailed than [Section 2](#), as recent ML-based systems have received less scholarly attention. A fuller account of their architectures and training processes is needed to support the later epistemic and ontological analysis. [Section 4](#) builds directly on the observations from both paradigms to trace how reductionist processes persist and to propose ways of ‘unleashing’ GenAI, and [Section 5](#) offers concluding reflections.

2 DEDUCTIVE LOGIC OF MUSIC GENERATION

Automatic music composition dates back to mediaeval times, as seen in the *Messe de Notre Dame* (1365) by Guillaume de Machaut. In the digital age, automatic music-generation systems typically operate within well-defined objectives and/or parameter spaces (e.g. algorithmic counterpoint, automatic accompaniment, exploration of possible variations). While many approaches have been proposed, we focus on those that define musical rules following classical symbolic AI methods, or ‘good old-fashioned AI’. In this approach, algorithms solve problems in a manner that resembles

human reasoning by representing problems as states and applying explicit logical rules to transition between these states. The approach is based on deductive logic: ‘intelligence’ is seen as a representation of the world that can be ‘formalised into propositions’ (Pasquinelli, 2023) and requires explicit symbolic representations and logical rules (Bajohr, 2024).

The creator of an algorithmic music-composition tool based on symbolic intelligence must manually code the compositional rules and logic that are typically derived from music theories and specific styles. For instance, Schottstaedt (1984) created a system to automatically compose music in the polyphonic style of Palestrina, using rules from a 1722 composition manual (Fux et al., 1971), while Ebcioğlu (1990) imitated Bach’s chorales by defining around 350 rules. Perhaps one of the most sophisticated of these systems was developed by Cope (1992), who combined defined rules with Markov chains to imitate the styles of different classical tonal composers. In other cases, composers defined their own rules based on original artistic ideas. For instance, Cage, in HPSCHD, defined rules to recombine music ideas based on existing works using random values (Austin et al., 1992). A key aspect of this process was identifying the fundamental musical structures and processes significant to the composer.

In all cases, rule-based music generation encounters a necessary process of progressive ontological reduction of ‘what is music’ or ‘what counts in music’ – what Born (2010) called *analytical ontology*. This reduction starts from what we might refer to as the ‘music potential’ – the theoretical and infinitely large set of all possible sound combinations that can be deemed as music, or, in other words, the *music* that underlies and makes possible all *musics* – and progressively reduces it by adding constraints.

These progressive reductions can also be seen as sacrificial decisions that act as exclusionary filters. In the following sections, we identify three sequential exclusionary processes.

2.1 CHOICES OF MUSICAL PARAMETERS

The first reduction entails identifying the specific set of musical parameters or processes that the composer aims to automate. Western composition has traditionally focused on manipulating certain parameters. These parameters have historically been centred on pitch and time, but, in the last century, other parameters have gained prominence, such as timbre and dynamics (e.g. Luigi Nono), or rhythmic and harmonic textures (e.g. György Ligeti). These parameters, generally speaking, depend on the objective of the musician. For instance, Cope (1992) sought to develop computer models that could replicate specific Western composers’ styles. The initial stage involved identifying the specific musical parameters that characterise a particular composer’s

style by analysing existing scores and identifying salient features, which can then be translated into computational rules. For instance, one algorithm examined an artist’s repertoire for similarities at the level of metre, tactus and mode (ibid.) – parameters central to Classic music (broadly, the period from Bach and Brahms) but less crucial in other periods.

Although the musician’s objectives described above are deliberate and intentional, other less intentional factors also contribute to filtering the *music potential*, i.e. the musician’s subjectivity and the cultural context. The musician’s subjectivity encompasses their musical preferences, experiences and perceptual and cognitive skills that define their musical predisposition and appreciation. The cultural context includes the dominant cultural, political and social paradigms and episteme. In particular, algorithmic imitations of Western tonal music prioritise specific parameters and compositional strategies over others. Notable examples are the counterpoints in Hiller and Isaacson (1979) and Schottstaedt (1984) and the focus on harmony in Ebcioğlu (1990).

2.2 MUSICAL GRAMMATISATION

The musical parameters that survived the previous selection need to undergo a process of *grammatisation*, which Stiegler (1998) describes as the process of transforming continuities into discrete elements – like writing, which breaks ‘into discrete elements the flux of speech’. This concept relates to that of *commensuration*: the process of transforming qualities into quantities (Espeland and Sauder, 2007) so that qualities can be numerically represented and then replaced (Husain, 2021). This logic, which is typical in machinic automation and in the capitalistic mode of production (Pasquinelli, 2023), entails a sacrificial process insofar as there is space only for a finite number of ‘possibilities of operation and usages’ (Simondon, 2011).

We term *musical grammatisation* the process of simplification and fragmentation required to break down music into a discrete and lower-dimensional space of manageable properties, as well as the reductionist process caused by the specific epistemic tools used to compose music. Notably, grammatisation predates algorithmic composition. In traditional (non-digital) music, performers control parameters like pitch, timbre and timing as continuous elements. However, when inscribed into scores, these parameters are imposed over a grid of *allowed* note pitches and values, becoming discredited, uniformised or simply not accounted for (for instance, variations on timing were largely present in the Classic period but were not always represented on paper). Thus, notation systems and music theory are forms of music grammatisation.

In the digital domain, extra layers of grammatisation are added. Digital music representations – whether raw audio or symbolic – are reductionist processes. Raw audio

directly encodes physical information about recorded sound and is influenced, albeit marginally, by the recording technology and processes. Symbolic representation encodes only a subset of structural parameters (e.g. pitch, modes, rhythm) or performance features (e.g. timing, dynamics, techniques). The Musical Instrument Digital Interface, the most commonly used representation (and communication) protocol, often praised for musical universality, in fact undergoes a process of reductionism and simplification as it reduces the infinitely complex space of music to a few discrete parameters (Morrison and McPherson, 2024).

Furthermore, the algorithms and models that are typically used to process digital music – the MIR toolbox – are exclusionary. While undeniably useful, rich and sophisticated, the MIR toolbox can indeed only account for aspects that can be extracted, analysed, quantified and synthesised using epistemic processes of Western Empiricism. For instance, aspects such as micro-timing deviations (e.g. rubato) or timbre adjustments introduced in performance are often unaccounted for, as they are more difficult to measure and quantify. Similarly, qualities imbued with culturally specific meanings may lack formal descriptors or even vocabulary, making them invisible to MIR tools and, thus, epistemically excluded.

Once a composer has defined the set of musical parameters and processes they want to control, lower-dimensioned their complexity and translated them into a computer-intelligible representation, they need to develop models to manipulate symbols (input) and then elaborate a solution (output) in a using a symbolic formalism. This process inherently involves a curatorial dimension, as decisions regarding the generative process must be made regardless of the generation strategies employed. Notably, these decisions are shaped not only by the composer's expertise but also by the available technological resources.

2.3 LESSENERED ONTOLOGY

At the end of the chain of reductionist processes, the *music potential* is narrowed down to a highly constrained set of possibilities, and this is a necessary condition for deductive approaches to music generation. The result is a lessened music ontology that relies on the methods and technologies that are available and endorsed at a certain time by a certain culture. Thus, the top-down understanding of what music should be, along with the tools and theories chosen to reach this understanding, defines what music is. Effectively, epistemology (our knowledge about music) determines ontology (what music is).

Also, since epistemic tools are entangled with their object, carrying epistemological and ontological consequences (Barad, 2007; Giraud, 2019), the materials used for grammatisation influence what music is and what music can be represented. This influence is shaped

by political forces as the space of possible *musics* is confined by Western canons and 'protected by boundaries that reflect Western aesthetics' (James, 2019). Even in traditional music, the epistemic influence of written scores led to overlooking performance nuances, favouring harmony-melody-based music. Meanwhile, timbre, which was difficult to reduce to numbers without computational analysis, was less explored by Western composers.

From a Simondonian perspective, musical grammatisation is about analysing or representing the musical phenomenon *after* it has been conceptually fragmented or constituted into discrete parts. Thus, arbitrarily establishing *a priori* specific elements and quantities might fail to grasp the dynamic and potential-rich process of *becoming*, which Simondon (2020) termed *individuation*. Individuation emerges from a metastable, pre-individual state rich with potentials. Musical grammatisation stabilises certain musical structures by resolving tensions within a metastable continuum of musical possibilities. Yet, in doing so, it sacrifices alternative configurations. Musical grammatisation thus determines the expressive vocabulary available to composers and, by consequence, determines the musical possibilities that can be automated using rule-based approaches to generation and those that are excluded and thus unexplored. As a result, the richness of music's *pre-individual continuum* can only partially actualised.

The deductive logic of music generation is thus a curatorial one, as it unavoidably includes reductionist processes and assumptions about music that significantly limit and bias the space of possible musical outcomes. In the next section, we explore the extent to which ML-based music generation overcomes this limitation.

3 INDUCTIVE LOGIC OF MUSIC GENERATION

Data-based AI paradigm, also called ML, connectionism or indexical AI (Weatherby and Justie, 2022), stems from neuroscientists' efforts to mimic brain functions in computational terms (McCulloch and Pitts, 1943). This paradigm holds that intelligence comes from the empirical experience of the world, which operates through approximation and optimisation and learns from examples (Pasquinelli, 2023). The underlying logic is *inductive*: patterns and structures found in data are projected into new contexts.

Below, we present a representative set of paradigms selected for their prominence in contemporary music-generation research and commercial systems, and because they, as we shall see, exemplify reductionist mechanisms, sacrificial processes and assumptions about music. We mostly focus on self-supervised (unsupervised) models, where patterns are learnt without

human input, as these might theoretically reduce arbitrary choices and biases. This section is organised to follow the generative pipeline: from training-stage levers to inference-stage paradigms and conditioning.

3.1 TRAINING

A GenAI system normally entails two stages: training and inference. During training, the model is optimised to abstract musical representations by encoding input data into a structured latent space. In the inference stage, the model decodes or samples from this latent space to generate new outputs, usually in response to user inputs.

3.1.1 Data Accumulation

GenAI requires the accumulation of data – not just musical data but also text, videos, captions and lyrics. Only a limited portion of music is accumulated in training datasets, resulting in limited and biased² datasets (Born, 2020; Holzapfel et al., 2018; Huang et al., 2021). Since the generated output is directly linked, although not linearly, to *what goes in* (recorded music), the space of possible outcomes is highly constrained, which has important implications in the context of our study. First, the possibility of being recorded applies only to musical works performed within the past 150 years and, even within this period, many works were never recorded. Second, training data are biased towards Western content and aesthetics (Tao et al., 2024). Third, recording technologies are not neutral, as they are affected by the same ontological and epistemological issues explained in Section 2.3.

Training data used in GenAI thus represents a partial and biased catalogue of music. Given the inherent impossibility of compiling domain-exhaustive training data and the direct dependence of GenAI outputs on training data, the model's 'understanding' is necessarily limited and partial. While this limitation is virtually inconsequential for many MIR tasks, we argue that a partial, selective and biased generation poses significant ethical and epistemological challenges.

Given that the paper has an onto-epistemological focus, ethical discussion is purposely limited. However, we proposed such discussions in depth elsewhere.³

3.1.2 Data representation and pre-processing

Music training data can be represented as waveforms or as symbolic notation. Symbolic notation is already encoded, and thus normally requires little to no pre-processing. By contrast, while some models like WaveNet (Oord et al., 2016) operate solely in the time domain, waveforms are generally converted into a representation that captures spectral characteristics. A widely used representation is the log Mel spectrogram, which normalises the frequency scale to align with human perception, making it particularly relevant for music-generation tasks prioritising perceptual quality (Ma et al., 2024).

From a reductionist perspective, this phenomenological representation may initially appear neutral. However, working in the frequency domain privileges harmonic features, often at the expense of non-harmonic elements that might be better captured in the time-domain (ibid.). This prioritisation significantly impacts music generation, shaping what the model emphasises and overlooks. Consequently, neural codecs, log Mel spectrograms and Short-time Fourier transform (STFT) should be recognised as epistemic constructs rather than neutral tools, as their representational assumptions condition the ways in which music can be perceived, analysed and understood.

Other forms of raw data processing exist, which are based on common MIR feature extractors. These forms target specific musical elements, such as melody, dynamics and rhythm (Wu et al., 2023), thereby enacting a form of sacrificial reduction. Chroma features, for instance, divide the octave into the 12 pitch classes of Western equal temperament, thus privileging its tonal framework. Their alignment with the equal-tempered scale makes them effective for identifying musical elements such as chords, key signatures and modulations and is used by models like MusicGen (Copet et al., 2024) to guide the generation process.

Many GenAIs employ tokenisation as a pre-processing step. In order to be processed by an ML model, music data –whether raw audio or symbolic –must be segmented into sequences, or tokens. The choice of tokenisation method, particularly the distinction between discrete and continuous tokenisation, significantly shapes the latent representation, directly impacting the model's capacity to analyse, interpret and generate music.

Discrete tokenisation is often preferred for autoregressive models (Section 3.1.3) and symbolic music (Liao et al., 2024). This method represents short segments of music as individual tokens, converting music data into a sequence of distinct symbols to form a finite vocabulary. Discrete tokenisation thus imposes a level of *abstraction* and *simplification* onto music data and can be incentivised to focus on higher-level distinct musical elements. These elements may be explicitly defined (for instance, by specifying aspects such as melody, harmony and rhythm) or left to be determined by the datasets and loss functions. Either way, the process entails a reductionist approach akin to the grammatisation process described above. *Continuous tokenization* is normally used in diffusion (Section 3.1.3) for tasks requiring detailed acoustic representation (e.g. audio generation and speech recognition). This method maps music data into a continuous vector space and seems less affected by reductionist processes.

3.1.3 Model training

Tokens are used to train the model, enabling it to learn abstract representations of the data. GenAIs typically rely

on two main paradigms for training: diffusion and autoregressive models.

Diffusion models simulate a forward and reverse diffusion process. In the forward step, Gaussian noise is incrementally added to the music data until it becomes random noise, effectively removing all musical information. The model simultaneously learns to reverse this process by predicting and removing the added noise step-by-step, effectively reconstructing the original data (see [Section 3.2.1](#)).

Autoregressive models learn a probability distribution over possible next events based on the preceding sequence. Models like Museformer (Yu et al., 2022) use fine- and coarse-grained attention mechanisms to handle both long music sequences and local structural details. Bar-level fine-grained attention captures local structural information (e.g. detailed note sequences and short-term patterns), while coarse-grained attention gathers broader contextual information (e.g. analyses other bars and considers the global structure).

Both paradigms embed various other models at different stages for different tasks. In the remainder of this section, we review some of the most common ones, focusing on their relevance to our research question.

Autoencoders and *variational autoencoders* are often employed to learn lower-dimensional representations of data in an unsupervised manner. The architecture includes an *encoder* that compresses input into a lower-dimensional latent space and a *decoder* that attempts to reconstruct it. The latent representation, learnt at the bottleneck of the autoencoder, encodes a compressed representation that contains less but more high-level information than the input audio. The encoder and decoder are jointly trained to extract meaningful features in order to minimise a *loss function*, which typically combines reconstruction error with additional constraints (e.g. a regularisation term in variational autoencoders). The learnt features depend on the loss function and the encoder-decoder architecture. For example, a loss function that focusses on timing and pitch may lead the (variational) autoencoder to capture micro-timing nuances and precise pitch intervals. Conversely, a function targeting higher-level structures, such as harmonic progressions and melodic contours, would guide the encoder to represent chord changes.

In models based on *contrastive learning*, semantic properties emerge as the model learns to maximise similarity between similar samples and minimise similarity between dissimilar ones. This training objective encourages the model to capture high-level discriminative information, leading to representations that should encode meaningful features. Contrastive learning is often multimodal, in which case pairs of positive-negative samples originate from different modalities. Text-music pairing models like CLAP (Elizalde et al., 2022), MuSCALL (Manco et al., 2022) and MuLan (Huang et al., 2022) integrate

audio and language description into a joint multimodal latent space, extracting semantic features that bridge the two modalities. This approach helps relieve the generative model from learning the connection between audio and human semantics on its own. Beyond language, other modalities like videos and images can also be used to identify semantic tokens.

Context awareness involves developing latent representations by analysing the context of processed tokens. *Masked modelling*, for instance, trains models to predict missing or masked segments within input data, supposedly capturing higher-level, abstract and perceptually relevant features (Ma et al., 2024). Contextual information also plays a role in Generative pre-trained transformer (GPT's) self-attention mechanisms (Vaswani et al., 2017) that are used in music-generation systems like MusicLM (Agostinelli et al., 2023). These mechanisms enable models to weigh different parts of the input sequence, capturing long-range dependencies and complex relationships.

Hierarchical models are trained to represent and generate music at multiple levels of abstraction, from fine-grained details to high-level semantic concepts (Liao et al., 2024). This approach supposedly enables models to capture both the acoustic nuances and the overall meaning of audio. For instance, AudioLM (Borsos et al., 2023) employs a three-stage hierarchical framework to learn audio representation. First, semantic tokens are computed autoregressively using w2v-BERT (Chung et al., 2021) to capture long-term temporal structure. Next, coarse acoustic tokens, generated via SoundStream (Zeghidour et al., 2022), encode broad acoustic properties while remaining conditioned on semantic tokens. Finally, fine acoustic tokens refine audio quality by encoding sub-ttle details.

Some models employ structured hierarchies to represent musical concepts and guide generation. For instance, a symbolic music-generation model was trained with a four-level hierarchy: 'form' (music key and phrases), 'reduced lead sheet' (reduced melody and simplified chords), 'lead sheet' (melody and chords) and piano accompaniment (Wang et al., 2024). The authors propose that abstract music concepts at higher levels are enabled by stylistic specifications at lower levels. For example, they describe that 'a lead sheet is an abstraction implying many possible ways to arrange the accompaniment that shares the same melodic and harmonic structure, while an instantiated accompaniment is one of the possible realisations showing the accompaniment structure in more detail'. This reasoning appears predicated on an unverified assumption concerning the fundamental nature of musical structure.

3.2 INFERENCE

At the inference stage, pre-trained models use learnt patterns and user prompts (e.g. text or audio) to generate music that should align with the prompts.

3.2.1 Synthesis paradigms

The generation process depends on the specific paradigm. In *diffusion*, the model iteratively denoises the signal to reconstruct data that resemble the original distribution by leveraging noise predictions. This process generates new samples that fit the training data distribution rather than simply reconstructing it. Most⁴ diffusion systems use U-Nets in these iterations to predict progressively less-noisy versions of the input. Being a convolutional neural network, the U-Nets first identify salient features through convolutional filters that respond to patterns and structures within the input data. Then, in the downsampling (pooling) phase, dominant features are emphasised from simple patterns (e.g. short-term rhythms)⁵ to more abstract features (e.g. musical motifs) in deeper layers. During downsampling, only the most prominent and recurring patterns – those deemed ‘salient’ by the model – survive. Crucially, these features are not necessarily akin to how humans might perceive music.

In a process mirroring the learning phase, during inference, *autoregressive models* generate music sequentially, one note or event at a time. Beginning with a seed or empty sequence, the model iteratively predicts the next note, conditioning on the previously generated sequence.

The type of main paradigm and architecture has an effect on the generated music. Autoregressive models offer flexibility for real-time control and interactive generation due to their sequential approach, while diffusion models better capture long-term dependencies and thus overall music structure and coherence.

3.2.2 Conditioning

Conditioning modules guide generation towards specific characteristics or features during both training and inference, making music more *controllable* by aligning it with predefined parameters through text prompts for genre, mood, instrumentation or other musical parameters. MIDINet (Yang et al., 2017) and MelodyDiffusion (Li and Sung, 2023) condition generation on specific chord sequences to influence the harmonic structure of the output. Harmonising melodies is also used as a conditioning approach that involves generating a chord progression that complements a given melody, ensuring both harmonic and rhythmic alignment (Zhao et al., 2024). These features can include basic musical parameters such as rhythm, time signature, pitch and chords, as well as more complex textual descriptions like ‘a sad song with syncopated rhythm’ (ibid.). Thus, generation modules can theoretically be conditioned on any explicitly definable musical elements.

Some form of conditioning is often essential in GenAI models using diffusion. To generate a new sample (rather than reconstructing originals), the reverse process typically requires conditioning on additional information.

Many models use text descriptions, encoding them into text embeddings via text encoders. These embeddings condition the U-Net during denoising, guiding music generation to align with the text. This conditioning often uses cross-attention mechanisms within the U-Net, allowing text embeddings to influence the denoising process at multiple points in the network.

4 DISCUSSION

In [Section 2](#), we showed that rule-based systems entail assumptions about music and curatorial and sacrificial reductionist processes caused by an interference of epistemology into ontology, ultimately disallowing certain *musics*. We also showed that these processes are intrinsic to the deductive logic of music generation and thus might only be solved with a shift to a different logic of music generation. [Section 3](#) investigated whether inductive logic-based systems solved this limitation. We surveyed the main processes and methods involved in these models with the specific goal of identifying the presence of reductionist processes and assumptions.

4.1 EPISTEMIC MEDIATION ON MUSIC ONTOLOGY

[Section 3](#) revealed multiple procedures that simultaneously foreclose musical possibilities and introduce semantic distortions.

The data-accumulation stage acts as an exclusionary process through major curatorial decisions, with the consequence that the latent representation is inevitably learnt without exhausting the domain. The weights of the neural network and the resulting correlations are thus local and partial, rather than universal. Indeed, ML abstractions have no connection to a universal, but rather rely on ‘infinite particulars created anew for each particular place and time’ (Joque, 2022). Borrowing from Desai et al. (2022), the ‘data ontology’ in which GenAI exists thus necessarily affects the space of possibilities. This foundational flaw inevitably affects all subsequent learning and generation stages.

Tokenisation is an example of grammatisation in ML-based systems. Discrete tokenisation, in particular, represents short segments as individual tokens, converting the data into a finite vocabulary of distinct symbols. This process imposes a level of abstraction and simplification onto the music data, favouring the focus on higher-level, distinct musical elements like melody or harmony. Other curatorial and exclusionary processes occur throughout training. Specific choices of representation, pre-processing and tokenisation methods, as well as the adoption of one paradigm over another and fine-tuning the loss-function on specific objectives, involve sacrificial processes that restrict the space of possible outcomes. Thus, the epistemic process by which a partial latent space is pooled and recombined to generate new

material has an effect on the music ontology, which ends up being *lessened*. This specific outcome thus mirrors the limitations found in deductive logic of generation.

Exclusionary processes are also operationalised through the deliberate alignment of generative models that are designed to favour specific outcomes. This operationalisation is largely enacted through *conditioning*. While conditioning improves generated music by aligning it with listeners' expectations and desired attributes (Huang et al., 2023; Schneider et al., 2023), it simultaneously enacts control over the generated music. Foucault's work on disciplinary power (Foucault, 1995) provides lenses and vocabulary to frame conditioning as an apparatus of normalisation that steers music generation towards certain paths and away from others. The process is *orthopaedic* in nature; it is an act of 'correcting deformities' by setting constraints that define what is deemed acceptable, correct or preferable and invalidating or disallowing deviations from these pre-set norms.

As with inductive logic, conditioning references musical theories and prioritises certain elements, hard-coding preferred outputs through limited categories and parameters that define musical saliency (Born, 2021). This curation inevitably shapes the musical landscape, potentially homogenising creative output and manipulating taste. This risk is compounded by the dataset-level issues discussed in Section 3.1, where prior work (Ma et al., 2024; Morreale et al., 2023) has shown how opaque, culturally narrow collection practices can limit diversity in generative systems.

GenAI is also tainted by unexamined assumptions embedded within it. For instance, the hierarchical models' distinction between acoustic and semantic aspects of audio positions these two modes in an orthogonal relationship, which appears to contrast with theories of holistic music perception where acoustic features contribute to our interpretation of meaning and emotion (Juslin and Sloboda, 2010).⁶ The separation between fine- and coarse-grained attention to deal with musical aspects seems to be equally affected by this issue. Further assumptions emerge from uncritically adopting models designed for other domains. Autoencoders, which reduce the dimensionality of original data and enable the 'emergence' of semantic meaning, have proven useful in the textual domain but are questionable in music. Dimensionality reduction inevitably leads to information loss and distortion, as per (Shannon, 1948) rate-distortion theory. The challenge with music generation is that we cannot precisely quantify the implications of this information loss on *music potential*.

4.2 THE LIMITS OF LANGUAGE AS A SEMANTIC MEDIATOR

In deductive models, musical understanding and meaning are explicitly encoded. By contrast, GenAI models need to borrow semantics from extra-musical domains,

notably language. Models like MusCALL (Manco et al., 2022) and Mulan (Huang et al., 2022) operate on the assumption that text provides a suitable semantic bridge for developing latent representations and, in turn, generating *meaningful* music. While widely adopted due to these models' apparent success, this assumption rarely faces scrutiny. Language was not adopted for its theoretical alignment with the domain at hand (music) but because these models had succeeded in text prediction and were readily available.

It remains to be seen whether these passages in alien domains function neutrally, i.e. without interfering with the mediation process, or instead distort or misrepresent musical concepts. Deriving musical meaning from language rests on a flawed methodology because language is not an all-encompassing epistemic tool, as highlighted by prominent scholars across disciplines. Heidegger (1927) highlighted the inability of language to disclose all aspects of existence, Derrida (1976) critiqued its capacity to fully encapsulate meaning and Chomsky (1957) noted its insufficiency in expressing the full breadth of human thought. Additionally, human perception, experience and meaning involve pre-linguistic and unconscious dimensions that escape linguistic representation (Kristeva, 1993; Merleau-Ponty, 1945). Consequently, forcing *reality* (in this case, music) into language and linguistic structures inevitably filters it through pre-existing criteria and epistemic frames, diminishing its richness and complexity (Campagna, 2018, 2021). These limitations are particularly evident since music, unlike language, is predominantly non-propositional – it does not necessarily convey specific statements or propositions. Music can engage pre-linguistic and unconscious levels and preserve complexity without reducing it to rigid concepts that fail to capture all facets of human experience and thought.

It is also problematic, and surprising, that, while these semantic bridges are central in GenAI, there is little to no engagement with the scholarship from their original fields. As ML emerged from Cognitive and Data Science rather than Semiotics, text and language are treated as generic data, stripped of their inherent functions and complexities (Joque, 2022), largely disregarding established linguistic theories. This observation demonstrates unsound methodology, potentially resulting in a bias against non-'hard sciences', and also leads ML researchers to inadvertently reinvent established Semiotics insights. For example, the principle of opposition theory from linguistic structuralism, which posits that signs (words) derive meaning through differentiation from other signs rather than intrinsic properties (Saussure, 1916), closely mirrors the core principles of contrastive learning, where meaning is generated differentially (Bajohr, 2024). Similarly, the *context* component of Jakobson (1960)'s language framework is akin to the attention mechanism and masked modelling used

in ML. These parallels raise critical questions about valuable insights that remain untapped in GenAI development through a lack of engagement with semiotics and related fields.

Ultimately, irrespective of their implementations, the semantic bridges used in GenAI inevitably lead to information distortion and loss. This limitation might be inherent to the logic and thus insurmountable, as these models rely on an idea of music that is partial, oversimplified and flattened to fit the epistemic tools at hand, not unlike the reductionist processes of deductive logic. Thus, it is difficult to agree that audio-language learning can provide the high-level abstraction needed to close the semantic gap ([van den Oord et al., 2013](#)) in MIR ([Manco et al., 2022](#)). Notably, other semantic bridges like videos and images are likely to carry similar criticalities as they still rely on language mediation. Large-scale datasets, for instance, rely on YouTube titles and descriptions ([Gemmeke et al., 2017](#)) and human captions ([Agostinelli et al., 2023](#)).

4.3 A CHAIN OF SIGNIFICATION

Many GenAI systems repurpose existing material, including both training data ([Morreale et al., 2023](#)) and neural networks [Morreale \(2025\)](#) originally developed for different applications, for uses beyond music generation. For example, Noise2Music ([Huang et al., 2023](#)) uses a T5 text encoder to transform prompts into embeddings for diffusion models and relies on MuLan for pseudo-labelling unlabelled music clips. Similarly, MusicLM obtains semantics by cascading different models, such as ‘acoustic tokens’ from SoundStream or ‘semantic tokens’ from a speech-based mask modelling framework (w2v-BERT).

These modular pipelines belong to longer scholarly lineages. The ‘cascading module’ strategy, typical in Engineering and Mathematics, in particular, is a divide-and-conquer approach that breaks complex problems into smaller modules, each passing output to the next rather than working directly toward the final product. As the chain grows, each layer’s output moves one step further away from the original goal, leading to a progressive abstraction. In semiotic terms, this creates a chain of signification where each abstraction functions as a *signifier* that drifts further from the original *signified* (music). Eventually, these abstractions resemble ‘floating signifiers’ that are no longer firmly anchored to a specific signifier works.

The ontological and epistemological effects of this chain of signification are non-negligible. First, by relying on submodels to ‘solve’ specific music tasks (e.g. encoding and decoding, tokenisation, contrastive learning), GenAI systems inherit and transmit partial musical understandings. Second, these epistemic processes fragment the original task of automatic music creation into non-musical modules. As discussed, whether this practice is lossless, from a musical perspective, remains to be

seen. What is certain is that the risk of misinterpretation or semantic drift increases, as does the likelihood of compounding errors. For instance, in autoregressive models, minor inaccuracies introduced early can propagate and intensify at later stages.

These limitations echo the exclusionary and reductionist mechanisms identified in [Section 2](#). Whether arising from explicit compositional grammars in rule-based systems or from architectural and training choices in ML-based systems, they similarly constrain the space of possible musical outcomes. The next section builds on this parallel to consider how such constraints might be loosened by reframing the aims of GenAI.

4.4 UNLEASHING GENAI

In the previous sections, we explained how reductionist processes and unexamined assumptions shape musical possibilities in both rule-based systems (through explicitly encoded grammars) and ML-based systems (through the cumulative effects of data selection, representation and modelling choices). Despite these issues, GenAI might hold significant potential for musical innovation, which, however, is currently curbed by efforts to constrain these models to recreate predefined, biased and partial musical schemas using dubious semantic bridges.

Unleashing GenAI requires understanding and appreciating its unique ways of abstracting musical knowledge. Many modern ML models, particularly deep neural networks, are fundamentally based on the automated discovery of abstraction. This capability to abstract does not necessarily – and quite likely does not – overlap with how humans abstract ([Alvarado, 2023](#); [Fazi, 2021](#)). As [Watson \(2023\)](#) notes, ‘abstraction algorithms may be sufficient to identify some essential properties, but essential properties are not necessarily identifiable via abstraction algorithms’. Applied to music, this comment suggests that musical features humans deem essential might not hold the same status for GenAI, and this is possibly the other way around, too. These machine-derived representations might not mirror human musical reality but instead indicate a *machinic* musical reality or perhaps a Baudrillardian *hyperreality* – a simulacrum detached from musical referents.

Referring to Borges’ famous story about the fictional 1:1 scale map, [Baudrillard \(1994\)](#) commented:

Today, abstraction is no longer that of the map, the double, the mirror or the concept. Simulation is no longer that of a territory, a referential being or a substance. It is the generation by models of a real[ness] without origin or reality: a hyperreal. The territory no longer precedes the map, nor does it survive it.

Unlike rule-based systems, the music reality of GenAI is not derived from human experience or grounded in

human referents. Instead, it emerges from model-driven networks whose inner models and outputs constitute a form of machinic hyperreality. While trained on music composed and performed by humans, these models ‘understand’ and generate music by defining a new, self-referential musical space. Offenhuber (2024) argues that once a dataset establishes its own reality, the specific ways in which the data relate to the real world recede into the background. While the logic of deductive generation presupposes a *commensurable* frame between human intelligence and artificial processes (see Section 2.2), machinic hyperreality and human reality remain fundamentally *incommensurable*: they cannot be measured against each other (Fazi, 2021). As machinic musical hyperreality might escape human-like understanding of music and musical semantics, constraining generation tasks to human-accessible meanings might be short-sighted.

5 CONCLUSION

In this paper, we investigated the extent to which GenAI models still depend on exclusionary logic and ontological assumptions about music. First, we showed that the deductive approach’s reductionism is maintained in the inductive one, albeit in a different way. Human intervention and correction remain necessary at various stages, as does the presence of theories of music. Second, while GenAI does not rely on empirical (in the phenomenological sense, i.e. the embodied, subjective and perceptual grasp of musical properties that guides human composers) and subjective understanding and appreciation of musical properties, its current attempts to find musical meaning by leveraging text is problematic as it presupposes the flawed assumption that language can act as a mediator. Third, like rule-based systems, the tools used for music generation and the inability to exhaust the domain during training constrain the possible musical output. While identifying excluded music is challenging, as GenAI can potentially create any music (but so can any Turing-complete programming language (McPherson and Lepri, 2020)), they are in practice constrained by their epistemic structure. In fact, this might not be a technical failure of GenAI but rather an ideological failure of assuming that music is something that can be ‘solved’, a notion increasingly implicit in MIR research agendas that position musical understanding as a task solvable through data-driven optimisation (Born, 2021; Holzapfel et al., 2018).

We also showed that assumptions in ML-based systems are no more or less constraining than those in rule-based systems, although they differ in kind. While rule-based systems impose explicit, top-down constraints derived from compositional grammars, ML systems embed assumptions through choices in training data, pre-processing, architecture and conditioning

mechanisms. These assumptions often operate implicitly and are thus harder to detect or critique, but they are no less impactful. Finally, we suggested steering clear of coercion and unleashing GenAI. This suggestion has both ethical and artistic motivations. On the ethical side, GenAI, in its current form, necessarily entezrs in competition with human musicians (Morreale, 2021). This is especially the case with GenAI models designed (or allowed) to create music ‘in the style of’, as is the case with mainstream systems. However, an unleashed GenAI would not act as a cheaper surrogate but can instead open new spaces for genuine musical innovation. Artistically, an unleashed GenAI would not aim to replicate existing practices, nor would it be able to, due to the incommensurability argument we presented above. Thus, this very divergence in the human and machinic musical realities holds the potential to foster genuine innovation within the musical domain.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their meticulous work and generous support, which have significantly strengthened this paper.

COMPETING INTERESTS

The authors have no competing interests to declare.

NOTES

1. With GenAI systems, we refer to data-driven end-to-end ML systems for music generation, including, but not limited to, text-to-music.
2. In this article, we refer to *bias* as the exclusions that result from dataset construction and modelling choices rather than the technical use in ML of *bias* as stylistic regularities.
3. Large-scale datasets are often compiled through opaque and inconsistently documented processes Morreale et al. (2023), shaped by narrow socio-cultural backgrounds that prioritise Western popular music and underrepresent other traditions. Such practices risk reinforcing cultural homogeneity; marginalising less-documented musical ecosystems; and embedding sectarian ontologies, epistemologies and values into generative systems, further constraining their creative possibilities (Ma et al., 2024, pp. 52–57).
4. Some diffusion-based models like Evans et al. (2024) use DiT (Diffusion Transformers) instead of U-Nets.
5. These are speculative examples of what these dominant features might be. However, identifying these specific features requires a thorough work on interpretability, which is currently virtually absent in the music domain.
6. It can be correctly argued that *acoustic vs semantic* and *coarse vs fine* might simply be unsophisticated ways to term machine representations that do not correspond to our typical understanding of these words. Thanks to the anonymous reviewer for this insightful suggestion.

AUTHOR AFFILIATIONS

Fabio Morreale  <https://orcid.org/0000-0002-4512-4897>
Sony AI, Barcelona, Spain

Marco A. Martinez-Ramirez
Sony AI, Tokyo, Japan

Raul Masu

Conservatorio 'F.A. Bonporti' di Trento e Riva del Garda, Italy;
CMA HKUST (GZ) Guangzhou, China

WeiHsiang Liao

Sony AI, Tokyo, Japan

Yuki Mitsufuji

Sony AI, New York, USA

REFERENCES

- Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C.** (2023). MusicLM: Generating music from text. *arXiv:2301.11325* [cs, eess].
- Alvarado, R.** (2023). AI as an epistemic technology. *Science and Engineering Ethics*, 29(5), 32.
- Austin, L., Cage, J., and Hiller, L.** (1992). An interview with John Cage and Lejaren Hiller. *Computer Music Journal*, 16(4), 15.
- Bajohr, H.** (2024). *Whoever Controls Language Models Controls Politics* (pp. 189–195). Walther König.
- Barad, K.** (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning* (p. 542). Duke University Press.
- Baudrillard, J.** (1994). *Simulacra and Simulation*. University of Michigan Press.
- Born, G.** (2010). For a relational musicology: Music and interdisciplinarity, beyond the practice turn: The 2007 Dent Medal Address. *Journal of the Royal Musical Association*, 135(2), 205–243.
- Born, G.** (2020). Diversifying MIR: Knowledge and real-world challenges, and new interdisciplinary futures. *Transactions of the International Society for Music Information Retrieval*, 3(1), 193–204.
- Born, G.** (2021). *Artificial Intelligence, Music Recommendation, and the Curation of Culture: A White Paper* (p. 27). Schwartz Reisman Institute for Technology and Society, University of Toronto.
- Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N.** (2023). AudioLM: A language modeling approach to audio generation. *arXiv:2209.03143* [cs].
- Campagna, F.** (2018). *Technic and Magic: The Reconstruction of Reality*. Bloomsbury.
- Campagna, F.** (2021). *Prophetic Culture* (pp. 1–280). Bloomsbury Publishing.
- Chomsky, N.** (1957). *Syntactic Structures*. Mouton & Co.
- Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y.** (2021). W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *arXiv:2108.06209* [cs].
- Cope, D.** (1992). A computer model of music composition. In *Machine Models of Music* (pp. 403–425).
- Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A.** (2024). Simple and controllable music generation. *arXiv:2306.05284* [cs, eess].
- Derrida, J.** (1976). *Of Grammatology* (1st American ed.). Johns Hopkins University Press.
- Desai, J., Watson, D., Wang, V., Taddeo, M., and Floridi, L.** (2022). The epistemological foundations of data science: A critical analysis. *SSRN Electronic Journal*.
- Ebcioğlu, K.** (1990). An expert system for harmonizing chorales in the style of J.S. Bach. *The Journal of Logic Programming*, 8(1–2), 145–185.
- Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H.** (2022). CLAP: Learning audio concepts from natural language supervision. *arXiv:2206.04769* [cs, eess].
- Espeland, W., and Sauder, M.** (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1–40.
- Evans, Z., Parker, J. D., Carr, C. J., Zukowski, Z., Taylor, J., and Pons, J.** (2024). Long-form music generation with latent diffusion. *arXiv:2404.10301* [cs, eess].
- Fazi, M. B.** (2021). Beyond human: Deep learning, explainability and representation. *Theory, Culture & Society*, 38(7–8), 55–77.
- Foucault, M.** (1995). *Discipline and Punish: The Birth of the Prison* (2nd Vintage Books ed.). Vintage Books.
- Fux, J. J., Mann, A., and Edmunds, J.** (1971). *The study of counterpoint from Johann Joseph Fux's Gradus ad Parnassum* (Rev. ed., 32nd printing). W. W. Norton.
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M.** (2017). Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. IEEE.
- Giraud, E. H.** (2019). *What Comes After Entanglement? Activism, Anthropocentrism, and an Ethics of Exclusion*. Duke University Press.
- Griffith, N., and Todd, P. M.** (1999). Frankensteinian methods for evolutionary music composition. In *Musical Networks*. The MIT Press.
- Heidegger, M.** (1927). *Being and Time*. Blackwell Publishing.
- Hiller, L., and Isaacson, L. M.** (1979). *Experimental Music: Composition with an Electronic Computer*. Greenwood Press.
- Holzapfel, A., Sturm, B. L., and Coeckelbergh, M.** (2018). Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval*, 1(1), 44–55.
- Huang, A., Sturm, B. L. T., and Holzapfel, A.** (2021). De-centering the west: East Asian philosophies and the ethics of applying artificial intelligence to music. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*, pp. 301–309.
- Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. W.** (2022). MuLan: A joint embedding of music audio and natural language. *arXiv:2208.12415* [cs, eess, stat].

- Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., and Han, W. (2023). Noise2Music: Text-conditioned music generation with diffusion models. *arXiv:2302.03917* [cs, eess].
- Husain, A. (2021). *Replace Me*. Peninsula Press.
- Jakobson, R. (1960). *Linguistics and Poetics*. Selected Writings III/Mouton.
- James, R. (2019). *The Sonic Episteme: Acoustic Resonance, Neoliberalism, and Biopolitics*. Duke University Press.
- Joque, J. (2022). *Revolutionary Mathematics*. Verso.
- Juslin, P. N., and Sloboda, J. A. (Eds.). (2010). *Handbook of Music and Emotion: Theory, Research, Applications* (p. xiv, 975). Oxford University Press.
- Kristeva, J. (1993). *Revolution in Poetic Language* (nachdr. ed.). Columbia University Press.
- Li, S., and Sung, Y. (2023). MelodyDiffusion: Chord-conditioned melody generation using a transformer-based diffusion model. *Mathematics*, 11(8), 1915.
- Liao, W., Takida, Y., Ikemiya, Y., Zhong, Z., Lai, C.-H., Fabbro, G., Shimada, K., Toyama, K., Cheuk, K., Martínez-Ramírez, M. A., Takahashi, S., Uhlich, S., Akama, T., Choi, W., Koyama, Y., and Mitsufuji, Y. (2024). Music foundation model as generic booster for music downstream tasks. *arXiv:2411.01135* [cs].
- Ma, Y., Øland, A., Ragni, A., Del Sette, B. M., Saitis, C., Donahue, C., Lin, C., Plachouras, C., Benetos, E., Quinton, E., Shatri, E., Morreale, F., Zhang, G., Fazekas, G., Xia, G., Zhang, H., Manco, I., Huang, J., Guinot, J., . . . Wang, Z. (2024). Foundation models for music: A survey. *arXiv:2408.14340* [cs, eess].
- Manco, I., Benetos, E., Quinton, E., and Fazekas, G. (2022). Contrastive audio-language learning for music. *arXiv:2208.12208* [cs].
- McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133.
- McPherson, A., and Lepri, G. (2020). *Beholden to our tools: Negotiating with technology while sketching digital instruments*. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, p. 6. Birmingham City University.
- Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Routledge.
- Morreale, F. (2021). Where does the buck stop? Ethical and political issues with AI in music creation. *Transactions of the International Society for Music Information Retrieval*, 4(1), 105–113.
- Morreale, F. (2025). Human subsumption in training datasets for music generation. In *The Inner World of AI*. Routledge.
- Morreale, F., Sharma, M., and Wei, I.-C. (2023). *Data collection in music generation training sets: A critical analysis*. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, Milan, Italy.
- Morrison, L., and McPherson, A. (2024). *Entangling entanglement: A diffractive dialogue on HCI and musical interactions*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17. ACM.
- Offenhuber, D. (2024). Shapes and frictions of synthetic data. *Big Data & Society*, 11(2), Article 20539517241249390.
- Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv:1609.03499* [cs].
- Pasquinelli, M. (2023). *The Eye of the Master: A Social History of Artificial Intelligence*. Verso.
- Saussure, F. d. (1916). *Cours de Linguistique Générale* (Éd. critique [Reprint of the 1916 edition]). Payot.
- Schneider, F., Kamal, O., Jin, Z., and Schölkopf, B. (2023). Moûsai: Text-to-music generation with long-context latent diffusion. *arXiv:2301.11757* [cs].
- Schottstaedt, B. (1984). *Automatic Species Counterpoint* (No. 19). CCRMA, Department of Music, Stanford University.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423.
- Simondon, G. (2011). On the mode of existence of technical objects. *Deleuze Studies*, 5(3), 407–424.
- Simondon, G. (2020). *Individuation in Light of Notions of Form and Information* (Vol. 1). University of Minnesota Press.
- Stiegler, B. (1998). *Technics and Time*. Stanford University Press.
- Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), 346.
- van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates, Inc.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, U., and Polosukhin, I. (2017). *Attention is all you need*. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.
- Wang, Z., Min, L., and Xia, G. (2024). Whole-song hierarchical generation of symbolic music using cascaded diffusion models. *arXiv:2405.09901* [cs].
- Watson, D. S. (2023). On the philosophy of unsupervised learning. *Philosophy & Technology*, 36(2), 28.
- Weatherby, L., and Justie, B. (2022). Indexical AI. *Critical Inquiry*, 48(2), 381–415.
- Wu, S.-L., Donahue, C., Watanabe, S., and Bryan, N. J. (2023). Music ControlNet: Multiple time-varying controls for music generation. *arXiv:2311.07069* [cs, eess].
- Yang, L.-C., Chou, S.-Y., and Yang, Y.-H. (2017). MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv:1703.10847* [cs].
- Yu, B., Lu, P., Wang, R., Hu, W., Tan, X., Ye, W., Zhang, S., Qin, T., and Liu, T.-Y. (2022). Museformer: Transformer with fine-

and coarse-grained attention for music generation. *arXiv:2210.10349* [cs].

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2022). SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 495–507.

Zhao, M., Zhong, Z., Mao, Z., Yang, S., Liao, W.-H., Takahashi, S., Wakaki, H., and Mitsufuji, Y. (2024). OpenMU: Your Swiss Army knife for music understanding. *arXiv:2410.15573* [cs].

TO CITE THIS ARTICLE:

Morreale, F., Martinez-Ramirez, M.A., Masu, R., Liao, W., & Mitsufuji, Y. (2025). Reductive, Exclusionary, Normalising: The Limits of Generative AI Music. *Transactions of the International Society for Music Information Retrieval*, 8(1), 300–312. DOI: <https://doi.org/10.5334/tismir.256>

Submitted: 13 February 2025 **Accepted:** 13 August 2025 **Published:** 4 September 2025

COPYRIGHT:

© 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Transactions of the International Society for Music Information Retrieval is a peer-reviewed open access journal published by Ubiquity Press.